Replicating Revealing the Dark Secrets of BERT on GPT2-small

Varchita Lalwani

1 Problem Definition

My chosen task involves replicating a study named ”Revealing the Dark Secrets of BERT” [KRRR19] using GPT-2. This study was conducted by Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky from the Department of Computer Science at the University of Massachusetts Lowell in 2019. It falls under the category of studying the learned features in language models, specifically in the subcategory of miscellaneous topics and it has difficulty category B. It is numbered as 9.61 in the list of problems.

The document examines each experiment and provides detailed results from BERT, as outlined in the original paper along with the results obtained from experiments conducted on GPT2-small.

2 General Idea

The research questions this study focuses on are:

1. What are the common attention patterns, how do they change during fine-tuning, and how does that impact the performance on a given task?
2. What linguistic knowledge is encoded in self-attention weights of the fine-tuned models and what portion of it comes from the pre-trained model?
3. How different are the self-attention patterns of different heads, and how important are they for a given task?

3 Experiment details

All fine-tuning experiments were conducted using a batch size of 32, 3 epochs, and the Adam optimizer with a learning rate of 1e-3. Due to resource and time constraints, datasets with large sets were limited to a maximum of 3000 data points, while validation datasets were capped at 500 data points.

In each experiment, self-attention weights were extracted for each head in every layer, as per the paper's methodology. This resulted in 2D float arrays of shape L × L, where L represents the length of an input sequence. These arrays, referred to as self-attention maps, were analyzed to determine which target tokens received the most attention as the input was processed token by token. These experiments were then utilized to analyze the model's processing of various linguistic information, including different parts of speech (such as nouns, pronouns, and verbs), syntactic roles (such as objects and subjects), semantic relations, and negation tokens. For the exploration of self-attention maps, 200 data points were considered.

As BERT is a bidirectional model, the attention maps are square, whereas for GPT, they form a lower left triangle. GPT does not utilize [CLS] and [SEP] tokens. The experiments employed the setting of ”prepend bos” as True, and the ”endoftext” token was considered as the sentence separator between two sentences.

4 Results Overview

Experiments over GPT2-small suggest that it is significantly over-parametrized same as BERT. This is supported by the discovery of repeated self-attention patterns in different heads. Unlike BERT GPT does not have vertical maps for where sentences separate.

The experiments demonstrate a notable improvement in performance for fine-tuned GPT models across all datasets. However, this improvement is accompanied by substantial weight transformations in many layers of the GPT model. While linguistic features are inherited from the pre-trained weights of BERT, the experiments indicate that numerous layers of GPT undergo significant weight updates for downstream tasks, unlike BERT. This suggests that GPT learns linguistic features necessary for downstream tasks in its early layers and progressively learns more complex features present in the sentence in later layers. Nevertheless, it's important to note that this doesn't imply that GPT's pre-trained weights neglect linguistic features such as nouns and pronouns because experiments reveal that they also allocate significant attention to these features for their auto-regressive tasks.

Paper found no evidence that attention patterns (BERT) that are mappable onto core frame-semantic relations actually improve BERT's performance. In the later sections it is seen that 2 out of 144 heads that seem to be "responsible" for these relations do not appear to be important in any of the GLUE tasks: disabling of either one does not lead to a drop of accuracy. This implies that fine-tuned BERT does not rely on this piece of semantic information and prioritizes other features instead. This type of exploration is done on a very limited dataset but experiments say with limited evidence that frame-semantic relations may be helpful up to a certain threshold unlike BERT. To say this with concrete evidence i.e, disabling heads and layers is yet to be done.

Overall, I observed that certain claims behave differently in GPT compared to BERT. This raises the question of why GPT undergoes such significant weight modifications during fine-tuning, and what specific information it learns during pre-training that differs from the requirements of downstream tasks. Additionally, despite revising the paper, results, and experiments, I still encountered the same outcomes, leading me to question whether there may be errors in my understanding or interpretation.

5 Datasets

1. **MRPC:** The MRPC dataset consists of sentence pairs sourced from news articles, where each pair is labeled to indicate whether the sentences are semantically equivalent (paraphrases) or not.

2. **STSB:** Semantic Textual Similarity Benchmark dataset. Each example consists of two sentences and a similarity score indicating the degree of similarity between the two sentences. The similarity score is typically a continuous value ranging from 0 to 5, where 0 indicates no similarity and 5 indicates complete similarity.

3. **SST-2:** Stanford Sentiment Treebank dataset. The dataset consists of sentences from movie reviews and their associated binary sentiment labels (positive or negative).

4. **QQP:** Quora Question Pairs dataset. It is a dataset composed of pairs of questions from the community question-answering website Quora. The dataset is labeled to indicate whether the question pairs are semantically equivalent (duplicate) or not.

5. **RTE:** Recognizing Textual Entailment dataset. Given a pair of text sequences (premise and hypothesis), the task is to classify whether the relationship between them is entailment, contradiction, or neutral. The labels include "entailment" (the hypothesis logically follows from the premise), "contradiction" (the hypothesis contradicts the premise), or "neutral" (there is no clear logical relationship between the two).

6. **QNLI:** Question-answering Natural Language Inference dataset. A dataset derived from the Stanford Question Answering Dataset (SQuAD). Each example consists of a question, a premise

(typically a sentence), and a label indicating whether the question is entailed by the premise, contradicted by the premise, or unrelated to the premise. This is closely related to the Recognizing Textual Entailment (RTE) task but focuses specifically on the question-answering context.

7. **MNLI:** Multi-Genre Natural Language Inference dataset. Each example consists of a premise (typically a sentence) and a hypothesis (another sentence), along with a label indicating the logical relationship between them. The labels include "entailment" (the hypothesis logically follows from the premise), "contradiction" (the hypothesis contradicts the premise), or "neutral" (there is no clear logical relationship between the two). It has multiple genres of text, including fiction, government, telephone conversations, and more.

6 Experiments conducted to address the above research questions.

6.1 Self-attention patterns

6.1.1 Tasks

As manual inspection of self-attention maps (Figure 1b) for both basic pre-trained and fine-tuned BERT models suggested that there is a limited set of self-attention map types that are repeatedly encoded across different heads.

1. Vertical: mainly corresponds to attention to special BERT tokens [CLS] and [SEP] which serve as delimiters between individual chunks of BERT's inputs;
2. Diagonal: formed by the attention to the previous/following tokens;
3. Vertical+Diagonal: a mix of the previous two types,
4. Block: intra-sentence attention for the tasks with two distinct sentences (such as, for example, RTE or MRPC),
5. Heterogeneous: highly variable depending on the specific input and cannot be characterized by a distinct structure.

GPT2 also has limited set of self-attention map types that are repeatedly encoded across different heads and slight different from BERT. I have used prepend_bos as True.

1. *Diagonal (D): formed by the attention to the previous/following tokens;*
2. *Block (B): Some heavy Vertical blocks of attentions in both sentences,*
3. *Overall (O): attention present overall irrespective of darkness (weights),*
4. *Majority/all (M1): Majority/all attention in part 1 (sentence-1)*

6.1.2 Methodology

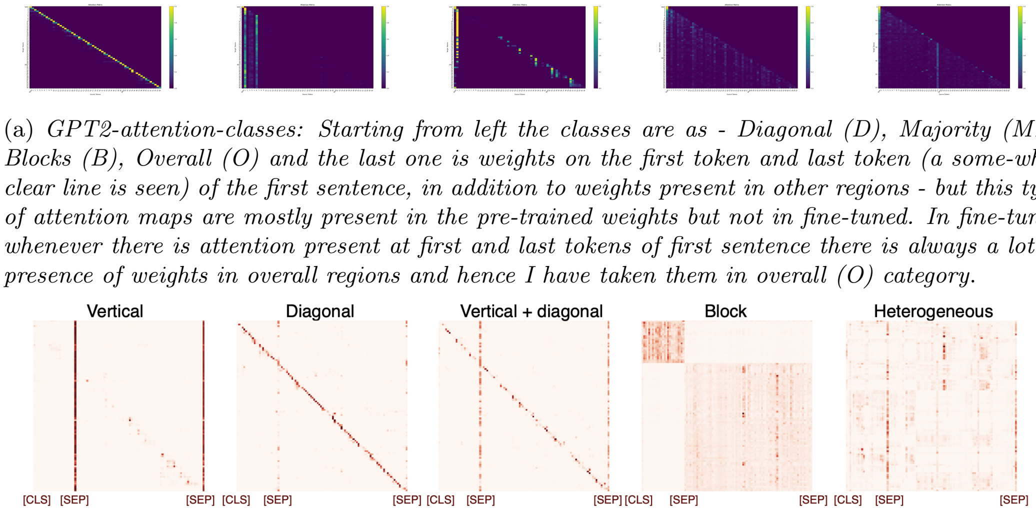
To get a rough estimate of the percentage of attention heads across different classes, paper has manually annotated around 400 sample self-attention maps as belonging to one of the five classes. The self-attention maps were obtained by feeding random input examples from selected tasks into the corresponding fine-tuned BERT model. This produced a some what unbalanced dataset, in which the "Vertical" class accounted for 30% of all samples. Then trained a convolutional neural network with 8 convolutional layers and ReLU activation functions to classify input maps into one of these classes. This also produced some what unbalanced dataset.

I annotated a total of 1294 samples of self-attention maps, assigning them to one of four classes. These annotations were generated by inputting random examples from selected tasks into the corresponding fine-tuned GPT2 model. Approximately 9% of the samples were classified as belonging to the "Blocks" class, 15% to the "Diagonal" class, 20% to the "M1" class, and 54% to the "Overall" class. However, this resulted in a somewhat unbalanced dataset. Then I attempted to classify the input maps into one of these classes using a model consisting of 8 convolutional layers, ReLU activation, stride as 2, and batch normalization.

6.1.3 Results

Paper achieved the F1 score of 0.86 on the annotated dataset. They then used this classifier to estimate the proportion of different self-attention patterns for the target GLUE tasks using up to 1000 examples from each validation set. While a large portion of encoded information corresponds to attention to the previous/following token, to the special tokens, or a mixture of the two (the first three classes), the estimated upper bound on all heads in the "Heterogeneous" category (i.e. the ones that could be informative) varies from 32% (MRPC) to 61% (QQP) depending on the task Figure 1b.

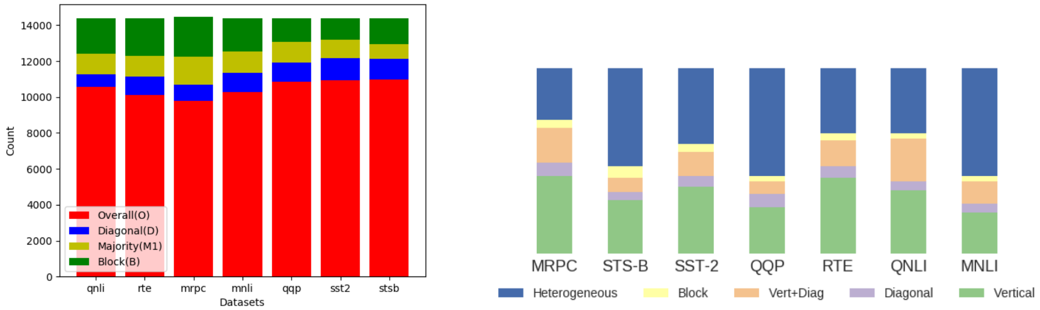
I attained an F1 score of 0.61 on the annotated dataset. Subsequently, I utilized this classifier to estimate the distribution of various self-attention patterns for the target GLUE tasks. I used up to 200 examples from each validation set for this analysis. The results are as in Figure 1a.



(a) *GPT2-attention-classes: Starting from left the classes are as - Diagonal (D), Majority (M1), Blocks (B), Overall (O) and the last one is weights on the first token and last token (a some-what clear line is seen) of the first sentence, in addition to weights present in other regions - but this type of attention maps are mostly present in the pre-trained weights but not in fine-tuned. In fine-tuned whenever there is attention present at first and last tokens of first sentence there is always a lot of presence of weights in overall regions and hence I have taken them in overall (O) category.*

(b) BERT-attention-classes: The first three types are most likely associated with language model pre-training, while the last two potentially encode semantic and syntactic information.

Figure 1: Typical self-attention classes used for training a neural network. Both axes on every image represent GPT2/BERT tokens of an input example, and colors denote absolute attention weights (darker colors stand for greater weights).



(a) *GPT2-attention-classes*

(b) BERT-attention-classes

Figure 2: Estimated percentages of the identified self-attention classes for each of the selected GLUE tasks.

6.2 Relation-specific heads in GPT2

6.2.1 Tasks

This section is to understand whether different syntactic and semantic relations are captured by self-attention patterns by selecting to examine semantic role relations defined in frame semantics, since they can be viewed as being at the intersection of syntax and semantics. Paper focused on whether BERT captures FrameNet's relations between frame evoking lexical units (predicates) and core frame elements and whether the links between them produce higher attention weights in certain specific heads. They used pre-trained BERT in these experiments. *I have used GPT2-small pre-trained in these experiments. FrameNet website is down so I couldn't get the data. I took 25 sentences and semantics from FrameNet site and performed this task.*

6.2.2 Methodology

As described in paper for each of these sentences, obtain pre-trained model's attention weights for each of the 144 heads. For every head, return the maximum absolute attention weight among those token pairs that correspond to the annotated semantic link contained within a given sentence. Then average the derived scores over all the collected examples. This strategy allows to identify the heads that prioritize the features correlated with frame-semantic relations within a sentence.

6.2.3 Results

For BERT the heatmap of averaged attention scores over all collected examples suggests that 2 out of 144 heads tend to attend to the parts of the sentence that FrameNet annotators identified as core elements of the same frame. The maximum attention weights averaged over all data examples for these identified heads account for 0.201 and 0.209, which are greater than a 99-th percentile of the distribution of values for all heads. Figure 3b shows an example of this attention pattern for these two heads. Both show high attention weight for "he" while processing "agitated" in the sentence "He was becoming agitated" (the frame "Emotion directed"). Paper interprets these results as limited evidence that certain types of linguistic relations may be captured by self-attention patterns in specialized BERT heads.
Due to limited number of examples, Figure 3a a lot of heads tend to attend to the part of the sentence that FrameNet annotators identified as core elements of the same frame but the attention weights are not significantly high. There are other frames present getting more weights. So these results can be considered as limited evidence that certain types of linguistic relations may be captured by self-attention patterns in specialized GPT heads. For sentence: 'I am late for the class', Figure 3a shows there is a link between I and late but but other tokens are getting more attention weights. At least 76 people were injured and buried in the rubble', results show that there is a link between people and buried but other tokens are getting more attention weights.

6.3 Change in self-attention patterns after fine-tuning

6.3.1 Tasks

Fine-tuning greatly impacts performance and to understand that paper has compared how attention changes for each task in the GLUE benchmark before and after fine-tuning.

6.3.2 Methodology

It measures this change using cosine similarity, which tells how similar the attention patterns are between the pre-trained and fine-tuned BERT models. Additionally, it also evaluates the contribution of the pre-trained BERT model to overall performance by comparing two weight initialization methods: using pre-trained BERT weights and randomly sampling weights from a normal distribution.

6.3.3 Results

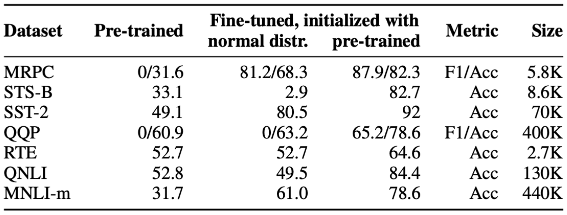
Figure 4b shows that for all the tasks except QQP, it is the last two layers that undergo the largest changes compared to the pre-trained BERT model. At the same time, Figure 2shows that finetuned

BERT outperforms pre-trained BERT by a significant margin on all the tasks. This leads to a conclusion that the last two layers encode task-specific features that are attributed to the gain of scores, while earlier layers capture more fundamental and low-level information used in finetuned models. BERT with weights initialized from normal distribution and further fine-tuned for a given task consistently produces lower scores than the ones achieved with pre-trained BERT. In fact, for some tasks initialization with random weights yields worse performance than pre-trained BERT without fine-tuning. This suggests that pre-trained BERT does indeed contain linguistic knowledge that is help ful for solving these GLUE tasks.

Table 4a shows that for all the tasks, all the layers except first two layers that undergo the largest changes compared to the pre-trained GPT2-small model. At the same time, **??***shows that finetuned GPT2 outperforms pre-trained GPT2 by a significant margin on all the tasks. The difference between BERT and GPT2 here is in GPT2-small only first 2 layers can be considered as earlier layers that capture more fundamental and low-level information used in fine-tuned models and remaining layers encode task-specific features that are attributed to the gain of scores while GPT2 with weights initialized from normal distribution and further fine-tuned for a given task consistently produces lower scores than the ones achieved with pre-trained GPT2-small. The results are not consistent with BERT as GPT2's all the layers except first two layers undergo the largest changes. This rises a question that whether GPT indeed contain linguistic knowledge or not that is helpful for solving these GLUE tasks like BERT. If not then what type of information is encoded in the GPT*

Dataset	**ND**	**PT**	**FT**
MRPC	0.38	0.59	0.65
STS-B	5.88	8.72	0.86
SST-2	0.5	0.61	0.75
QQP	0.57	0.51	0.67
RTE	0.45	0.46	0.58
QNLI	0.2	0.49	0.53
MNLI-m	0.25	0.267	0.36

Table 1: How well GPT2 models perform on different tasks in the GLUE benchmark. I have used 500 data points from the validation set. I have measured accuracy for most tasks, but for the STS-B dataset, we measured loss. The first column lists the datasets tested on. The next three columns show different ways of initialization of the models: ND means initialization with a normal distribution, PT means keeping all the pre-trained weights frozen and only training the classification head, and Init. PT initialization and fine-tuning the models with pre-trained weights.



Dataset	**Pre-trained**	**Fine-tuned, initialized with** **normal distr.**	 **pre-trained**	**Metric**	**Size**
MRPC	0/31.6	81.2/68.3	87.9/82.3	F1/Acc	5.8K
STS-B	33.1	2.9	82.7	Acc	8.6K
SST-2	49.1	80.5	92	Acc	70K
QQP	0/60.9	0/63.2	65.2/78.6	F1/Acc	400K
RTE	52.7	52.7	64.6	Acc	2.7K
QNLI	52.8	49.5	84.4	Acc	130K
MNLI-m	31.7	61.0	78.6	Acc	440K

Table 2: GLUE task performance of BERT models with different initialization.

6.4 Attention to linguistic features

6.4.1 Tasks

In this experiment, paper investigates whether fine-tuning BERT for a given task creates self-attention patterns which emphasize specific linguistic features. In this case, certain kinds of tokens may get high attention weights from all the other tokens in the sentence, producing vertical stripes on the corresponding attention maps. To test this hypothesis paper checked whether there are vertical stripe patterns corresponding to certain linguistically interpretable features, and to what extent such features are relevant for solving a given task. In particular, we investigated attention to nouns, verbs, pronouns, subjects, objects, and negation words, and special BERT tokens across the tasks.

6.4.2 Methology

For every head, paper computed the sum of self-attention weights assigned to the token of interest from each input token. Since the weights depend on the number of tokens in the input sequence, this sum is normalized by sequence length. This allows to aggregate the weights for this feature across different examples. If there are multiple tokens of the same type (e.g. several nouns or negations), take the maximum value. Disregard input sentences that do not contain a given feature. For each investigated feature, calculate this aggregated attention score for each head in every layer and build a map in order to detect the heads potentially responsible for this feature. Then compare the obtained maps to the ones derived using the pre-trained BERT model. This comparison enables to determine if a particular feature is important for a specific task and whether it contributes to some tasks more than to others.

6.4.3 Results

Contrary to paper's initial hypothesis that the vertical attention pattern may be motivated by linguistically meaningful features, Figure 5 they found that it is associated predominantly, if not exclusively, with attention to [CLS] and [SEP] tokens (Note that the absolute [SEP] weights for the SST-2 sentiment analysis task are greater than for other tasks, which is explained by the fact that there is only one sentence in the model inputs, i.e. only one [SEP] token instead of two. There is also a clear tendency for earlier layers to pay attention to [CLS] and for later layers to [SEP], and this trend is consistent across all the tasks. They did detect heads that paid increased (compared to the pre-trained BERT) attention to nouns and direct objects of the main predicates (on the MRPC, RTE and QQP tasks), and negation tokens (on the QNLI task), but the attention weights of such tokens were negligible compared to [CLS] and [SEP]. Therefore, they believe that the striped attention maps generally come from BERT pre-training tasks rather than from task-specific linguistic reasoning.

As Figure 7 shows a huge attention is paid to nouns, pronouns and verbs in pre-trained weights. The weight increase along the layers and all heads are contributing to weights. In fine-tuned weights, in the initial layers a lot of weights are heavy on nouns, pronouns and verbs but as we forward in layers the weights decrease. The higher layers goes under a lot changes implies focusing on task specific characteristics. So, whatever features passed down from pre-trained model are not directly for nouns or pronouns or etc. Also heavy change in weights can be the reason for a lot of of overall feature maps in self-attention patterns, as it learn relations from different parts.

6.5 Token-to-token attention

6.5.1 Tasks

To complement the experiments in attention to linguistic features and relation-specific heads, in this section, paper investigate the attention patterns between tokens in the same sentence, i.e. whether any of the tokens are particularly important while a given token is being processed. They were interested specifically in the verb-subject relation and the noun-pronoun relation. Also, since BERT uses the representation of the [CLS] token in the last layer to make the prediction, they used the features from the experiment in attention to linguistic features in order to check if they get higher attention weights while the model is processing the [CLS] token. The researchers aim to investigate how attention is distributed between tokens within the same sentence. They focus on identifying if certain tokens, such as verbs and subjects or nouns and pronouns, receive higher attention weights while processing other tokens.

6.5.2 Methodology

I have done exploration for noun-pronoun relationship. I extracted sentences in which both noun and pronoun appear and extracted their weights as described in the above section.

6.5.3 Results

Token-to-token attention experiments for detecting heads that prioritize noun-pronoun and verb-subject links resulted in a set of potential head candidates that coincided with diagonally structured attention maps. They believe that this happened due to the inherent property of English syntax where the dependent elements frequently appear close to each other, so it is difficult to distinguish such relations from the previous/following token attention coming from language model pre-training. Their investigation of attention distribution for the [CLS] token in the output layer suggests that for most tasks, with the exception of STS-B, RTE and QNLI, the [SEP] gets attended the most, as shown in Figure 6. Based on manual inspection, for the mentioned remaining tasks, the greatest attention weights correspond to the punctuation tokens, which are in a sense similar to [SEP].

In GPT, (Figure 8) we can see token-to-token attention experiments for detecting heads that prioritize noun-pronoun link resulted in a set of potential head candidates that coincided with diagonally structured attention maps. As above we can say that this happened due to the inherent property of English syntax where the dependent elements frequently appear close to each other, so it is difficult to distinguish such relations from the previous/following token attention coming from language model pre-training. However this directly hasn't come from pre-trained weights.

6.6 Disabling self-attention heads

6.6.1 Tasks

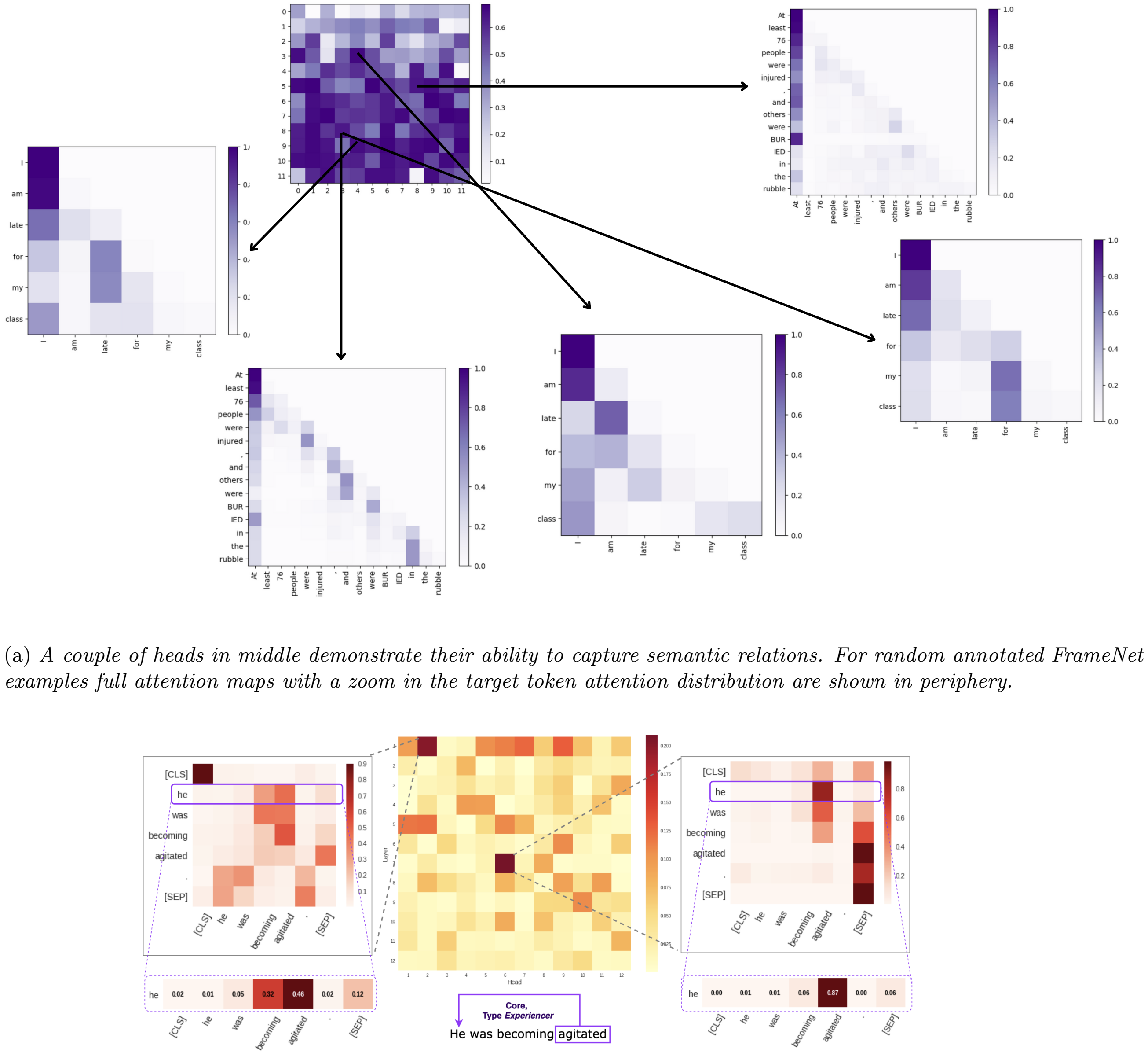
Since there does seem to be a certain degree of specialization for different heads, paper investigated the effects of disabling different heads in BERT and the resulting effects on task performance. They define disabling a head as modifying the attention values of a head to be constant a = L1 for every token in the input sentence, where L is the length of the sentence. Thus, every token receives the same attention, effectively disabling the learned attention patterns while maintaining the information flow of the original model.

6.6.2 Methodology

I got same results before and after disabling heads, I can sort this out in a little more time.

References

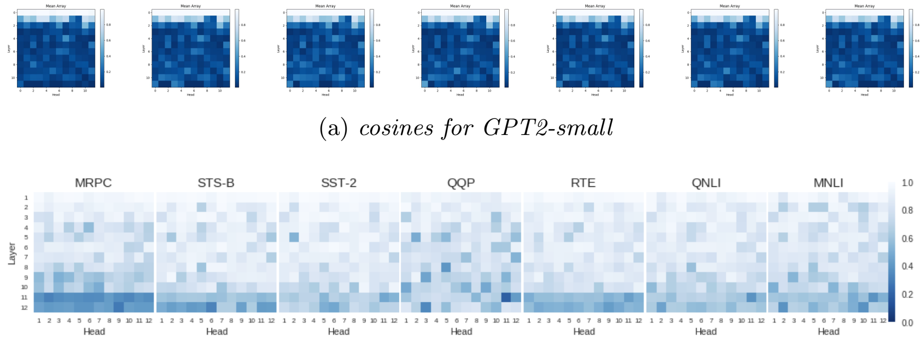
[KRRR19] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019.



(a) *A couple of heads in middle demonstrate their ability to capture semantic relations. For random annotated FrameNet examples full attention maps with a zoom in the target token attention distribution are shown in periphery.*

(b) Two heads (middle) demonstrate their ability to capture semantic relations. For one random annotated FrameNet example (bottom) full attention maps with a zoom in the target token attention distribution are shown (leftmost and rightmost)

Figure 3: Detection of pre-trained heads that encode information correlated to semantic links in the input text. Note that the middle heatmaps in the middle in both BERT and GPT2 are obtained through averaging of all the individual input example maps.



(a) *cosines for GPT2-small*

(b) cosines for BERT's

Figure 4: Per-head cosine similarity between pre-trained models and fine-tuned models' self-attention maps for each of the selected GLUE tasks, averaged over validation dataset examples. Darker colors correspond to greater differences. Starting from the left the datasets are: MRPC, STSb, SST2, QQP, RTE, QNLI, MNLI

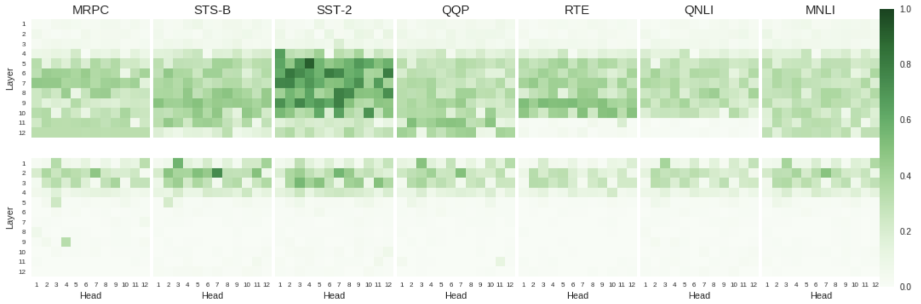


Figure 5: Per-task attention weights to the [SEP] (top row) and the [CLS] (bottom row) tokens averaged over input sequences' lengths and over dataset examples. Darker colors correspond to greater absolute weights.

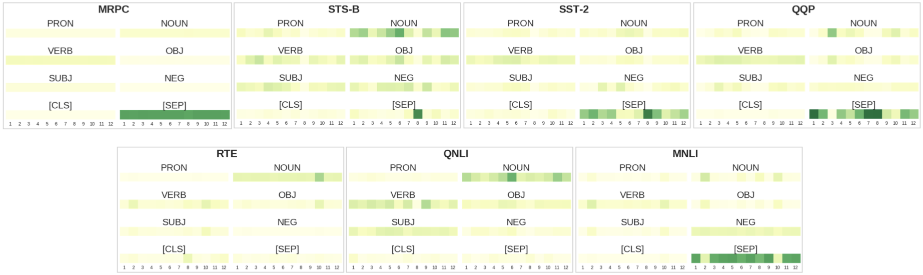
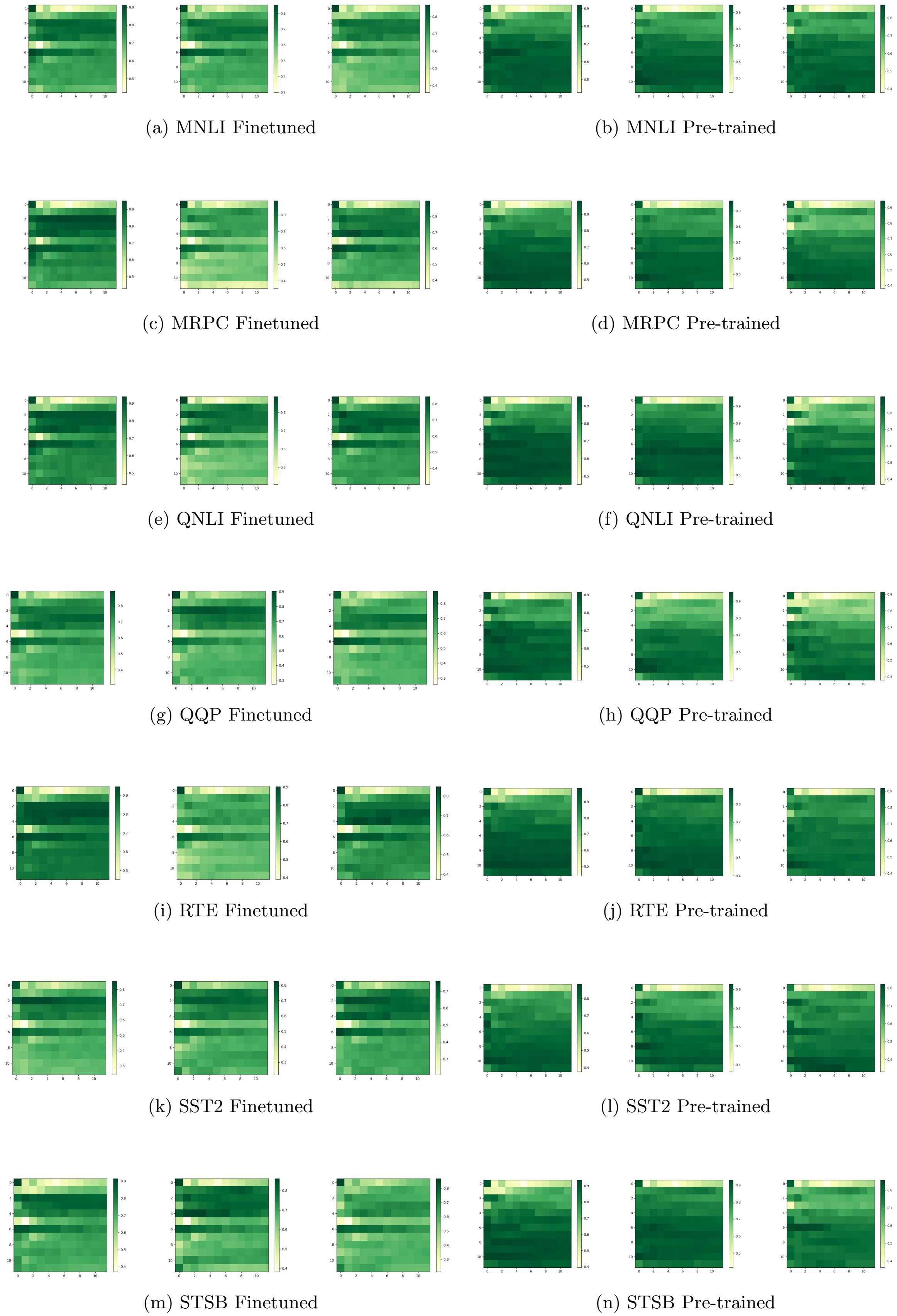


Figure 6: Per-task attention weights corresponding to the [CLS] token averaged over input sequences' lengths and over dataset examples, and extracted from the final layer. Darker colors correspond to greater absolute weights.



(a) MNLI Finetuned

(b) MNLI Pre-trained

(c) MRPC Finetuned

(d) MRPC Pre-trained

(e) QNLI Finetuned

(f) QNLI Pre-trained

(g) QQP Finetuned

(h) QQP Pre-trained

(i) RTE Finetuned

(j) RTE Pre-trained

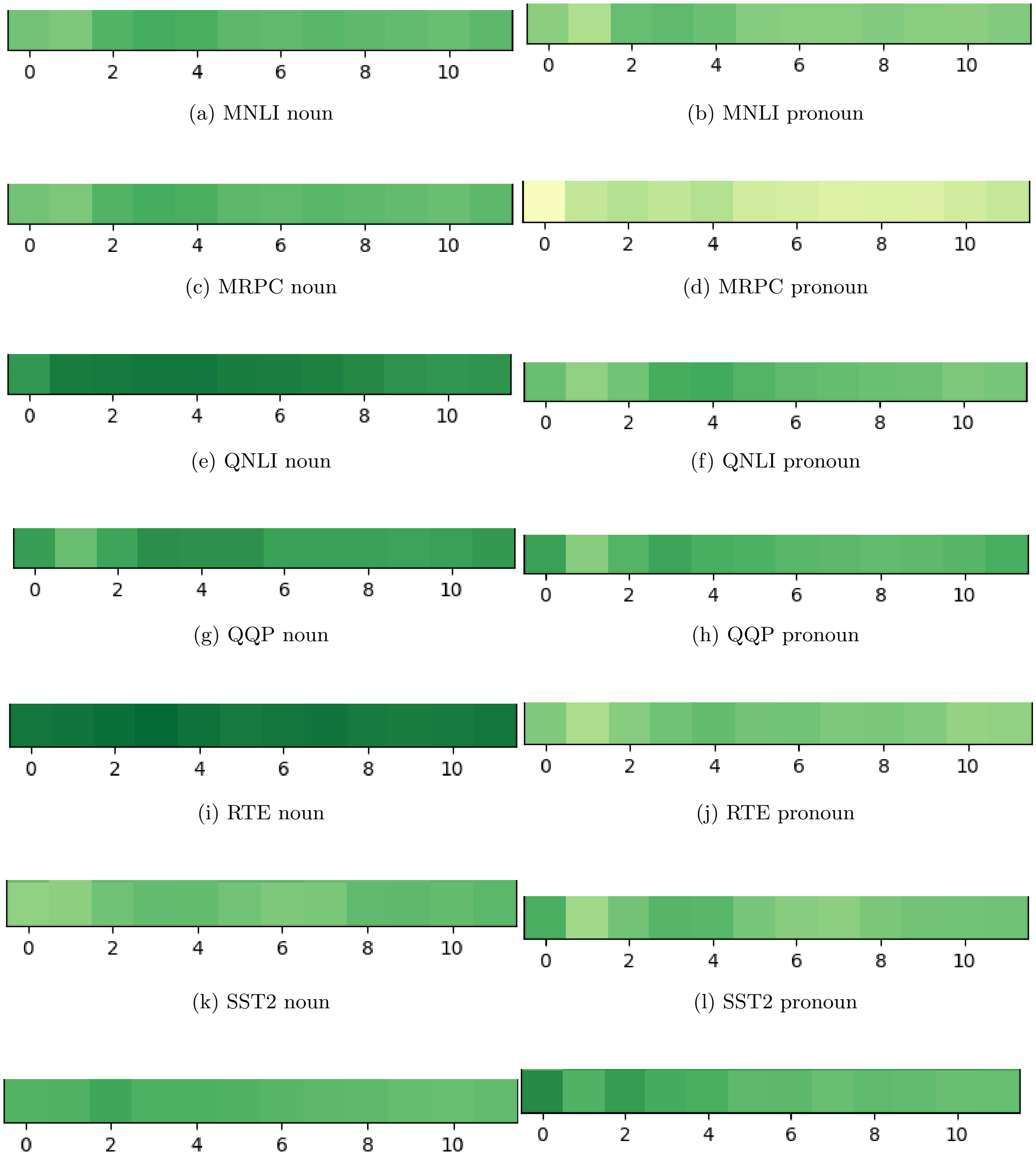
(k) SST2 Finetuned

(l) SST2 Pre-trained

(m) STSB Finetuned

(n) STSB Pre-trained

Figure 7: Per-task attention weights for Noun Pronoun and Verb tokens averaged over input sequences' lengths and over dataset examples. Darker colors correspond to greater absolute weights.



(a) MNLI noun

(b) MNLI pronoun

(c) MRPC noun

(d) MRPC pronoun

(e) QNLI noun

(f) QNLI pronoun

(g) QQP noun

(h) QQP pronoun

(i) RTE noun

(j) RTE pronoun

(k) SST2 noun

(l) SST2 pronoun

(m) STSB noun

(n) STSB pronoun

Figure 8: Per-task attention weights corresponding to the averaged over input sequences' lengths and over dataset examples, and extracted from the final layer. Darker colors correspond to greater absolute weights. The relation explored - noun and pronoun.